Building Computational Clusters

Session 615

Building Computational Clusters

Douglas Brooks
Product Manager, Server Hardware
d.brooks@apple.com

“Clustering”

Introduction

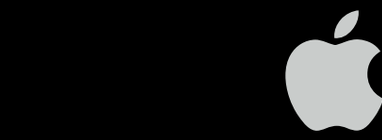
- What is “clustering”?
  - Clustering for high availability
  - Clustering for computation
- This session will be focusing on computational clusters
- Clustering for high availability will be touched on in “Deploying Xserve,” Session 621

What You'll Learn

- Why you might want a cluster
- Xserve and Mac OS X Server use in a cluster
- Typical cluster architecture
- Deploying an application on a cluster
- Physical and network concerns
- Tools and techniques for deploying a cluster
- Resources available for clustering

Why Clusters?

- Problem complexity > single CPU
- Computation need growing faster than CPU performance
- Large SMP boxes expensive and complex
- Easy to scale cluster to budget and problem
- Example applications
  - Image Manipulation (Rendering, Compositing)
  - Simulation (Biology, Products, Financial models)
  - Genomic analysis



Example

Rendering Pixar's Toy Story 2

6 hours/frame

24 frames/second

60 seconds/minute

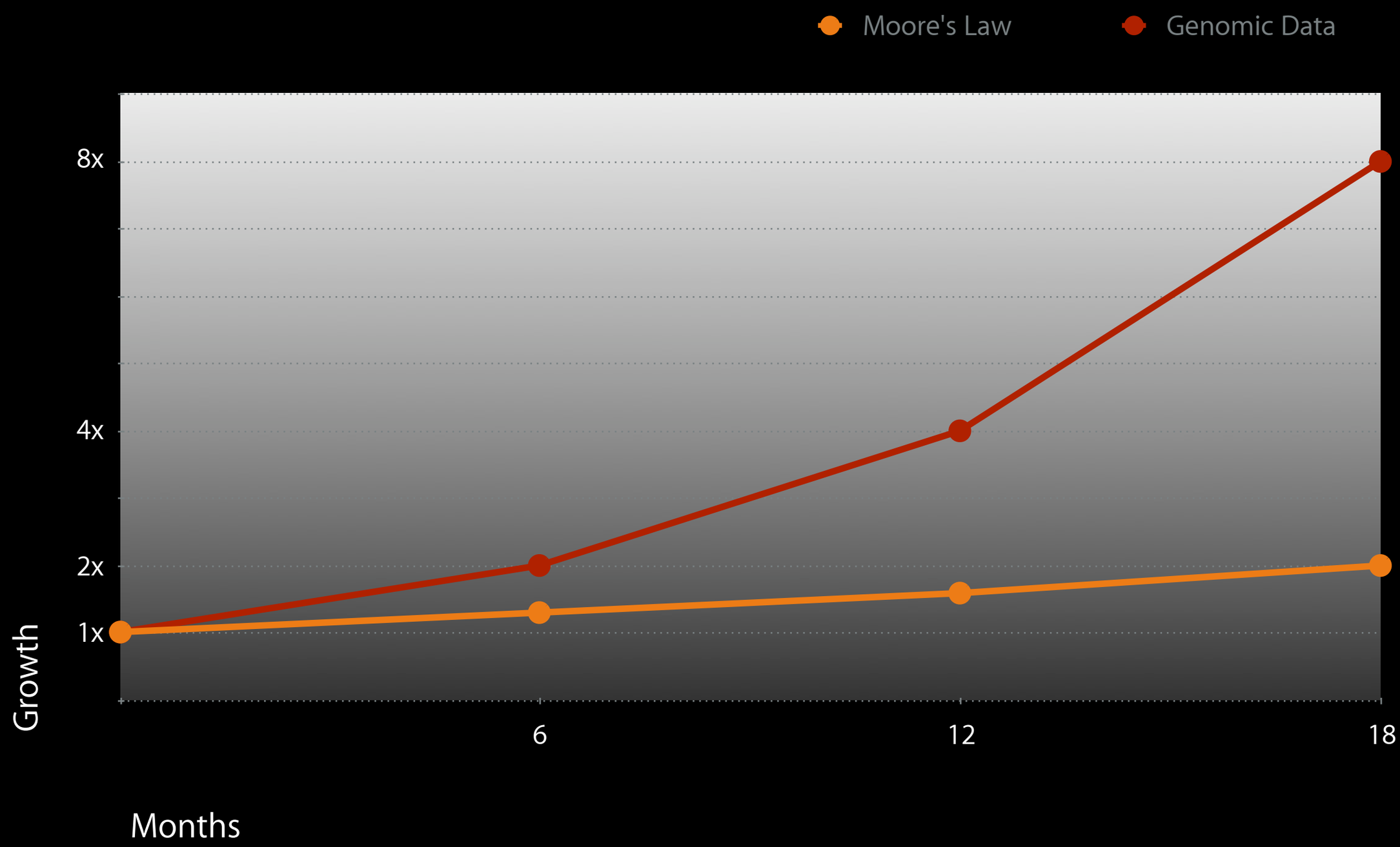
92 minutes

794,880 hours

92 years

Faster than Moore's Law

Growth in Genomic Data



Why Xserve?

- High-density processing power
- Velocity Engine
- Open source, BSD core UNIX operating system
- Easy to deploy and maintain
- Sophisticated remote management

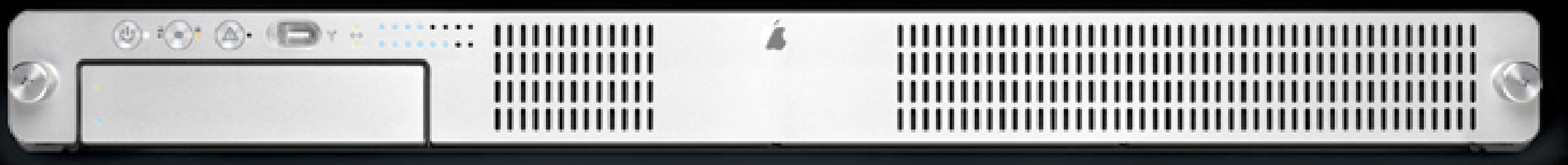
Why Xserve?

- High-density processing power
- Velocity Engine
- Open source, BSD core UNIX operating system
- Easy to deploy and maintain
- Sophisticated remote management
- Configuration optimized for compute tasks

Xserve Configurations

1 or 2 processors
4 drive bays–up to 720GB storage
Dual gigabit Ethernet

CD-ROM or combo Drive
VGA Video
Mac OS X Server Unlimited



2 processors
1 drive bay with 60GB ADM
On-board gigabit Ethernet

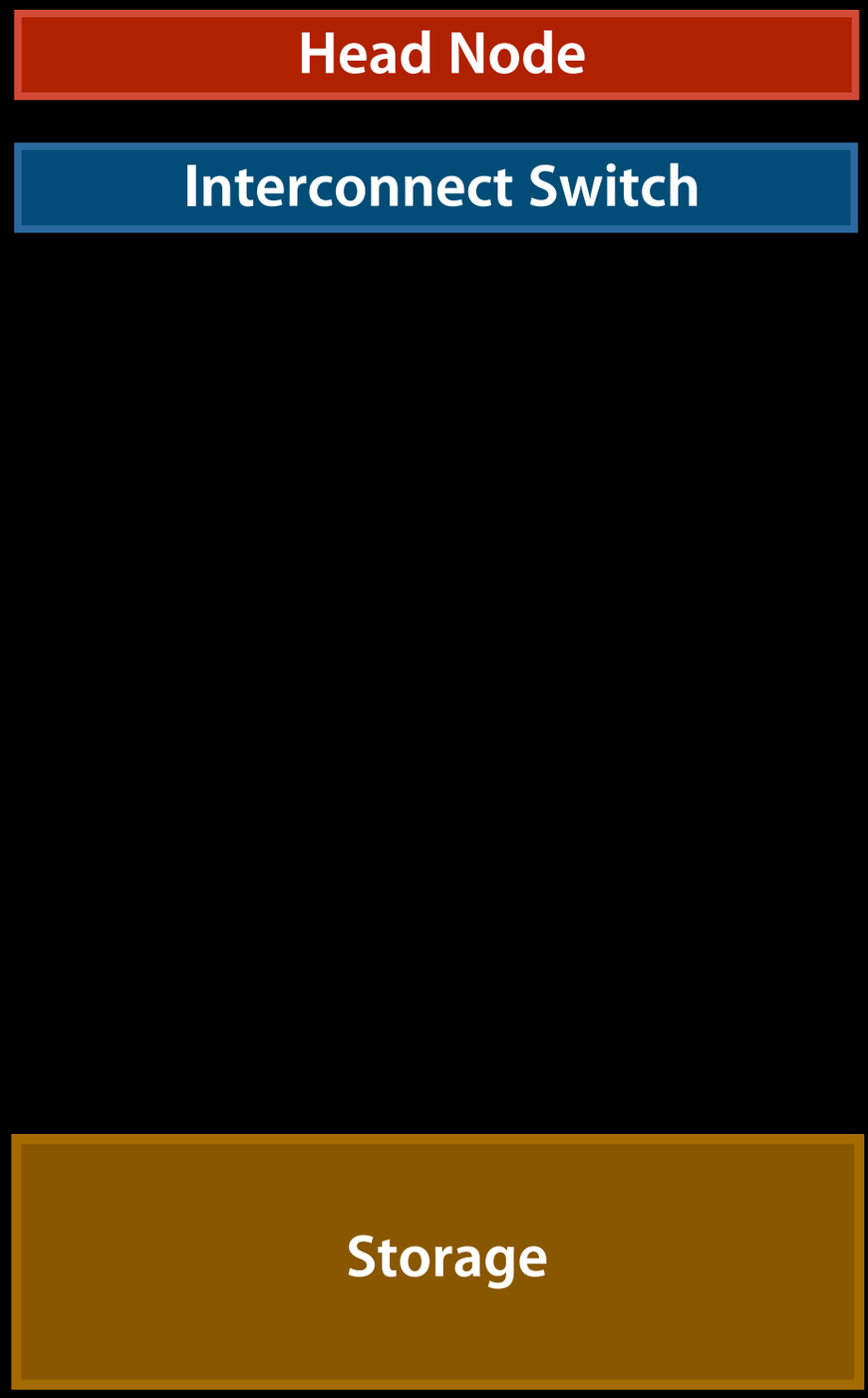
No optical drive
No video
Mac OS X Server 10-user license

Typical Cluster Architecture

- Head Node
  - Managed cluster resources
  - Manages user access
  - Provides centralized storage to cluster

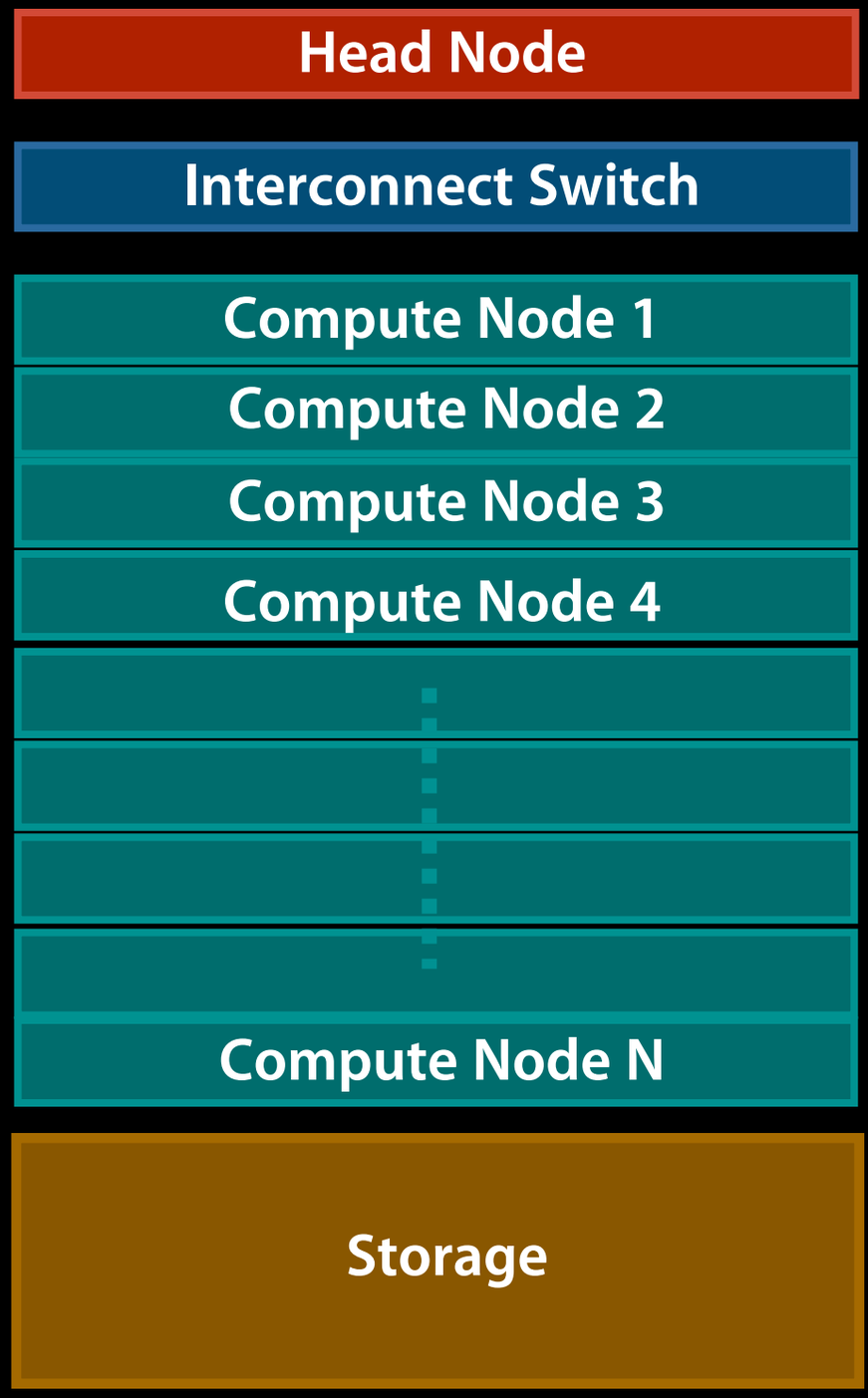
Typical Cluster Architecture

- Head Node
  - Managed cluster resources
  - Manages user access
  - Provides centralized storage to cluster
- Interconnect Network
  - Ethernet: 100 Mb/s or Gigabit
  - MyriNet
  - FireWire



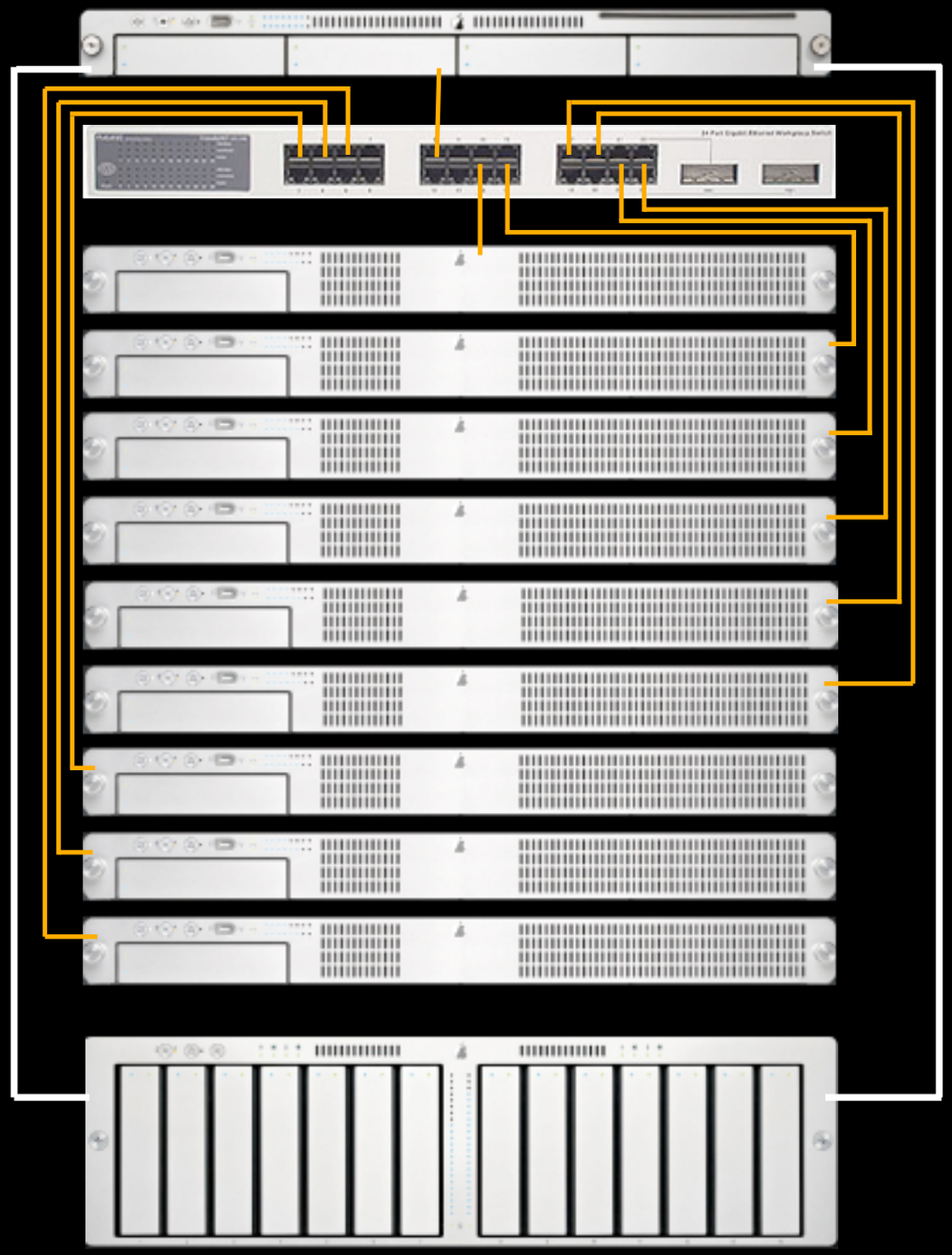
Typical Cluster Architecture

- Head Node
  - Managed cluster resources
  - Manages user access
  - Provides centralized storage to cluster
- Interconnect Network
  - Ethernet: 100 Mb/s or Gigabit
  - MyriNet
  - FireWire
- Compute Elements



Typical Cluster Architecture

- Head Node
  - Managed cluster resources
  - Manages user access
  - Provides centralized storage to cluster
- Interconnect Network
  - Ethernet: 100 Mb/s or Gigabit
  - MyriNet
  - FireWire
- Compute Elements



Deployment and Management

- Physical Concerns
  - Power and environmentals
  - Networking
- Logical Concerns
  - Software installation
  - Configuration
  - Remote management
  - Backup

Physical Concerns

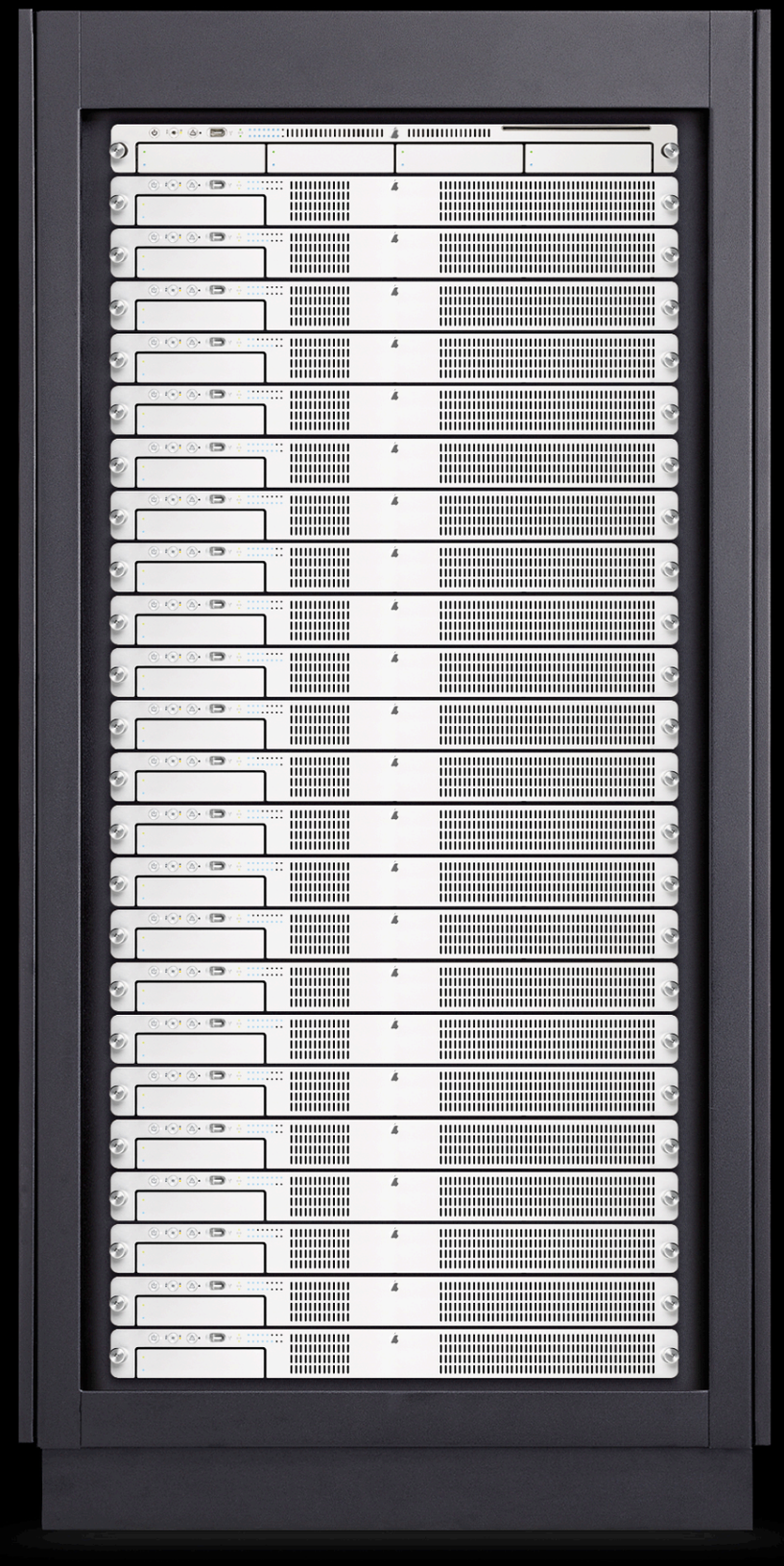
Power and Environmentals

- Xserve Dual Processor
  - Rated at 3.6A @ 345W for fully loaded system
  - **Typically 2A @ 134W**
  - 459 BTU/h

Physical Concerns

Power and Environmentals

- 16-node Cluster
  - Rated nearly 60A @ 5520W
  - Typically 32A @ 2200W
  - 7400 BTU/h
- Conclusions
  - Plan for requirements
  - UPS for head node and storage
  - Stagger startup



Physical Concerns

Network

- I/O bound or processor bound problem?
  - 100Mb vs. Gigabit
- Latency sensitive?
  - Ethernet vs. MyriNet
- Connectivity to outside world
  - Security
  - Access

Logical Concerns

- Software Installation
  - Cloning tools, NetBoot, Network Install
- Remote Management
  - ssh, web tools
- User Accessibility
  - Terminal access
  - Web interface
- Backup

Cluster Resources for Mac OS X

- Every major DRM package available
- Application availability for Mac OS X
  - Open source
  - Java
  - UNIX

Cluster Tools for Mac OS X

Distributed Resource Management

Platform LSF

http://www.platform.com

Sun Grid Engine

http://gridengine.sourceforge.net

http://www.bioteam.net/MacOSX/gridengine/

OpenPBS

http://www.openpbs.org

PBS Pro

http://www.altair.com/software/pbspro.htm

Cluster Tools for Mac OS X

High Performance Computing Tools

MPICH

http://www-unix.mcs.anl.gov/mpi/mpich/

MPI Pro

http://www.mpi-softtech.com/products/mpi_pro/macosx/

LAM/MPI

http://www.lam-mpi.org/

Linda and Paradise

http://www.lindaspaces.com/products/index.html

Pooch

http://www.daugerresearch.com/pooch/

Cluster Tools for Mac OS X

Grid Computing Tools

GridIron XLR8

http://www.gridironsoftware.com

Globus Toolkit

http://www.bioteam.net/MacOSX/

Cluster Tools for Mac OS X

Science and Research Applications

gridMathematica

http://www.wolfram.com/products/gridmathematica/

TurboBLAST, TurboHub, and TurboBench

http://www.turboworx.com

Bioinformatics Toolkit

http://www.bioteam.net/MacOSX/

iNquiry Bioinformatics

http://www.bioteam.net/

Accessible Clustering

Strategies for Scalable Computing

Michael Athanas, Ph.D.
Principal Investigator, Founding Partner
The BioTeam
michael@bioteam.net

http://bioteam.net

Introduction

Clustering enables boundless computing

- Pressures in life science computing
- Computing solutions
- Challenges in clustering
- Instant clustering

The BioTeam

The BioTeam

- Objective and vendor neutral informatics and “bio-IT” consulting
- Comprised of scientists experienced in bridging life science informatics and high performance IT
- The BioTeam principals:
  - Michael Athanas
  - Chris Dagdigian
  - Bill Van Etten
  - Stan Gloss

The BioTeam

The BioTeam

Current customers include:

- Apple Computer (USA, Canada, Asia)
- Agilix Corporation
- Beyond Genomics
- Dana Farber Cancer Institute
- Harvard University, Bauer Center for Genomics Research
- Michigan Center for Biological Information
- Pfizer
- Texas A&M University
- Third Wave Technologies
- Vertex Pharmaceuticals
- Wyeth Research

The BioTeam

Accessible Computing

Boundless computing

- Scientific research without constraints
  - Systems modeling and simulation
  - Data mining
  - Knowledge presentation

The BioTeam

Abstracting Computing Resources

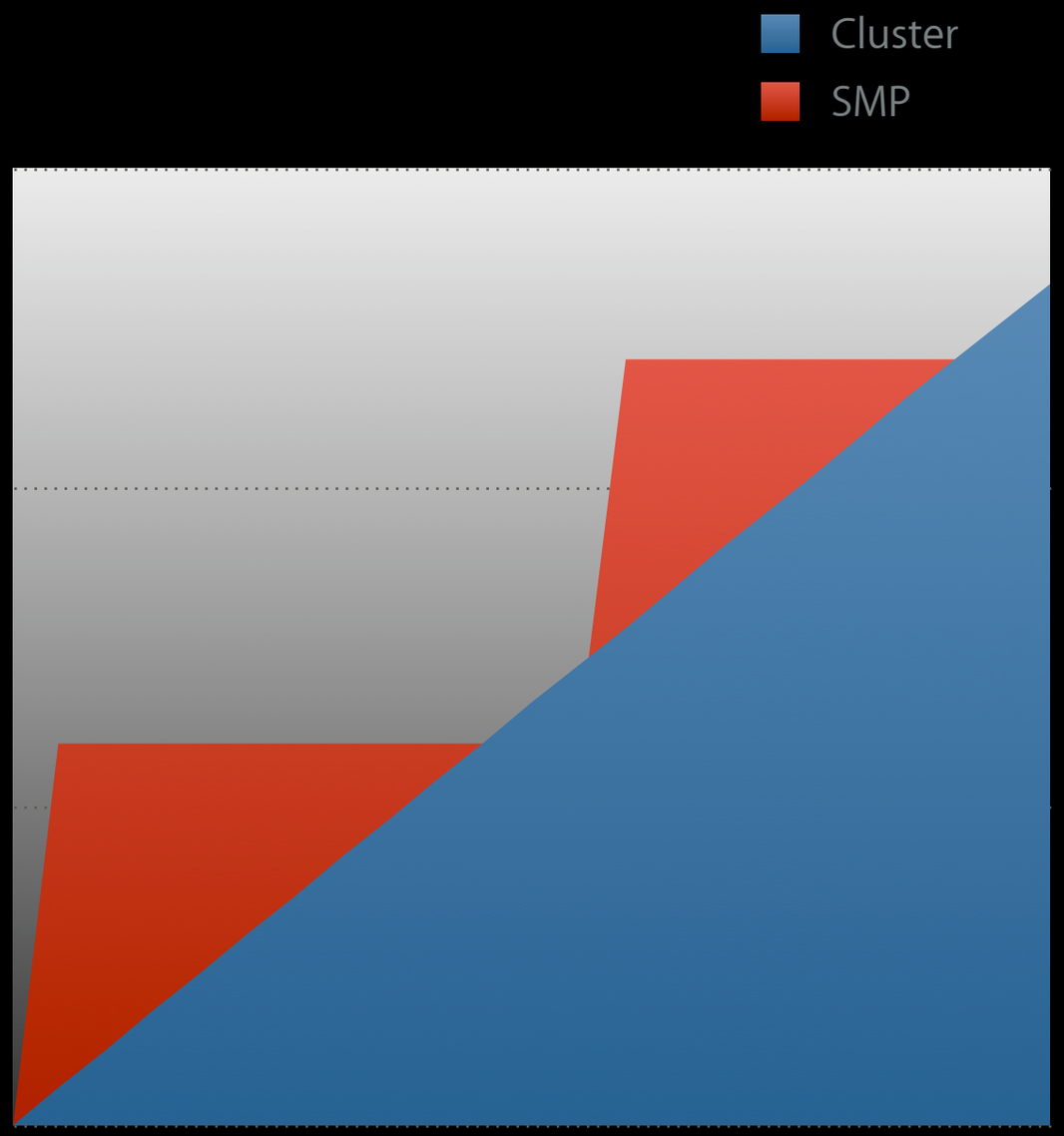
It's not about the computers, it's about applications and pipelines

What's Important	What's Not Important
Quick Data Access Reliable, Fast Execution Application Interoperability	ATA vs. SCSI RAID 5 vs. RAID 0 AMD vs. Intel vs. PowerPC

The BioTeam

Why Clustering?
Scalability

- Blueprint for growth
  - Flexible
  - Scalable
  - Incremental growth of capacity by “small” steps
  - Full capacity utilization
  - Unify distributed resources



The BioTeam

Why Clustering?

Price/Performance

- Commodity market server hardware is absolutely amazing
  - Clustering is an ideal way to harness that power

Why Clustering?
Flexibility

- Clustering enables application pipeline driven architectures:
  - Networking (low latency, high throughput)
  - Storage access (local cache, NAS, SAN, hybrid)
  - Processors (accelerators, SMP)
- Customized solutions
  - Transcend the “single-vendor solution”

The BioTeam

Why Clustering?
Reliability

- With careful architecture
  - Identification of single points of failure
    - Hardware and software
  - Minimal downtime compared to SMP
    - Compute elements are disposable

The BioTeam

Why Not Clustering?

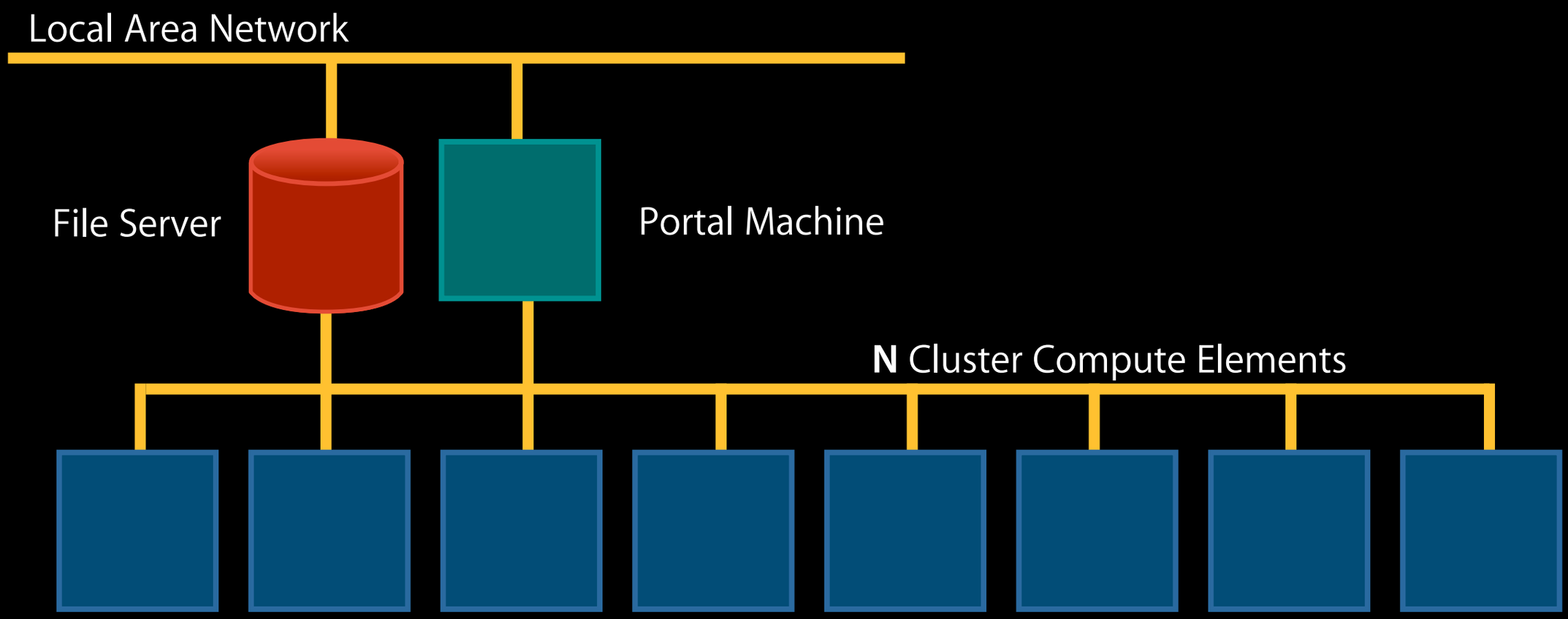
- Not all applications map to loosely coupled architectures
- Management complexity
  - Effort cannot scale with number of nodes
- User application complexity
  - Data parallel, PVM, MPI, DRM, …
- Reliability
- Achieving high utilization
  - High throughput and high performance

The BioTeam

The Portal Architecture

Abstracting computing resources

- Single point of entry
- Compute elements are generally inaccessible



The BioTeam

Cluster Maintenance Philosophy

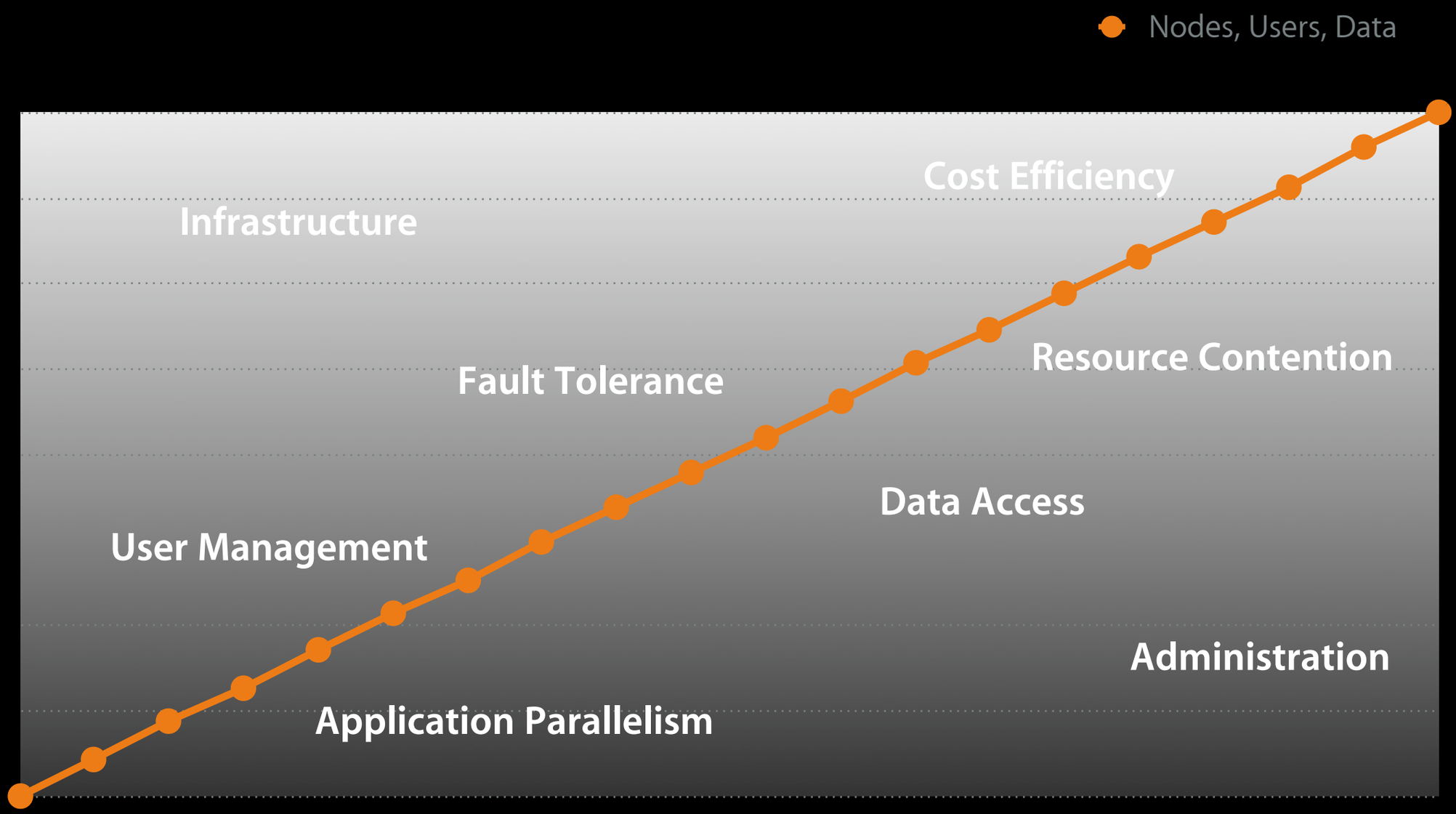
- Cluster compute elements must be
  - Anonymous
  - Interchangeable
  - Disposable
- Automation and remote management is key
  - Don’t waste admin resources messing around with individual servers and configuration settings
- Three possible states
  - Running, faulted, failed

The BioTeam

Cluster Maintenance Philosophy

- Prepare for high initial failure rates
  - Regardless of vendor
- Identify potential support contract and warranty issues
- Some questions to consider:
  - Debug or wipe OS problems?
  - Diagnose hardware or swap out?
  - Replace unit or components?
  - Internal staff or outsource?

Delivering Scalability



The BioTeam

Computing Strategies

Monolithic
Mainframe

Distributed
Cluster

Increasing application complexity

Increasing upfront cost

Increasing scalability

Increasing management TCO

The BioTeam

The BioTeam iNquiry

Instant Scalable Informatics!

Just Add Hardware

The BioTeam

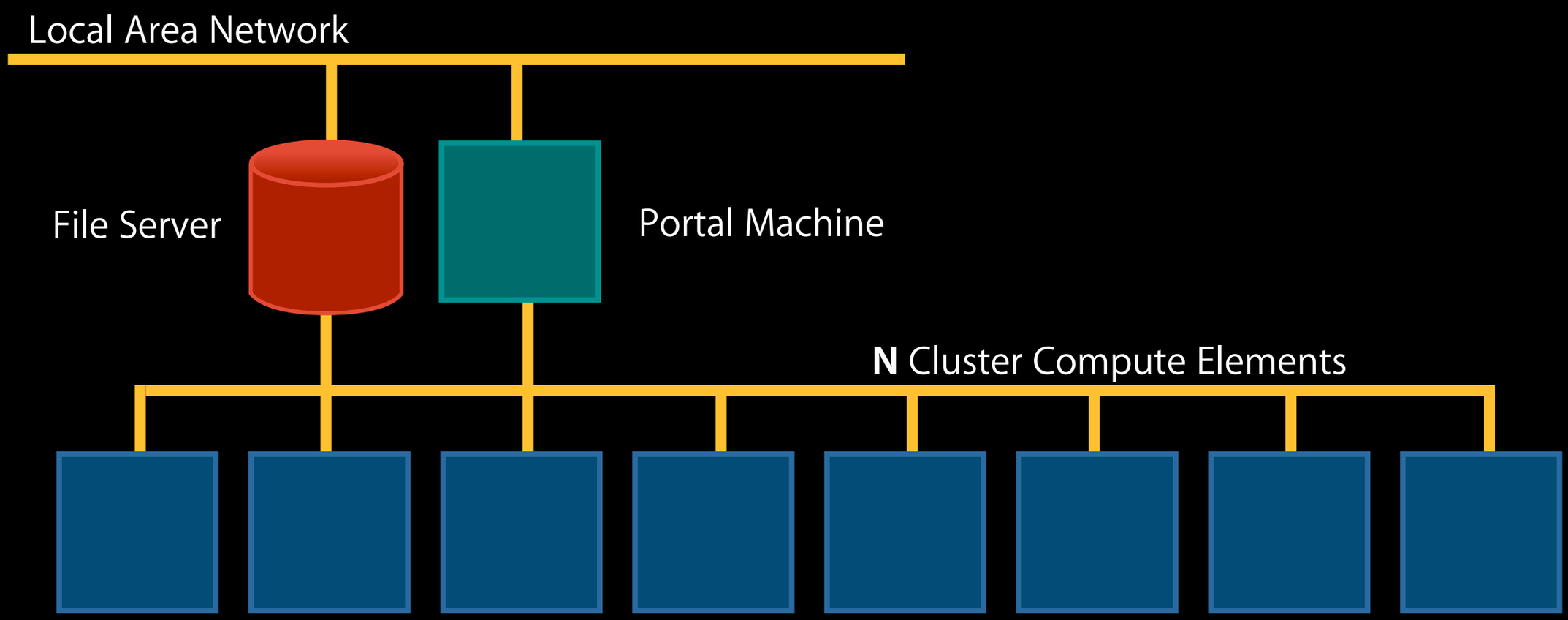
The BioTeam iNquiry

What's inside?

- Network services
- DRM (Sun Grid Engine)
- Administration and monitoring tools
- 200+ life science informatics applications
  - Cluster-enabled
  - User interface for the scientist
  - Consistent and extensible

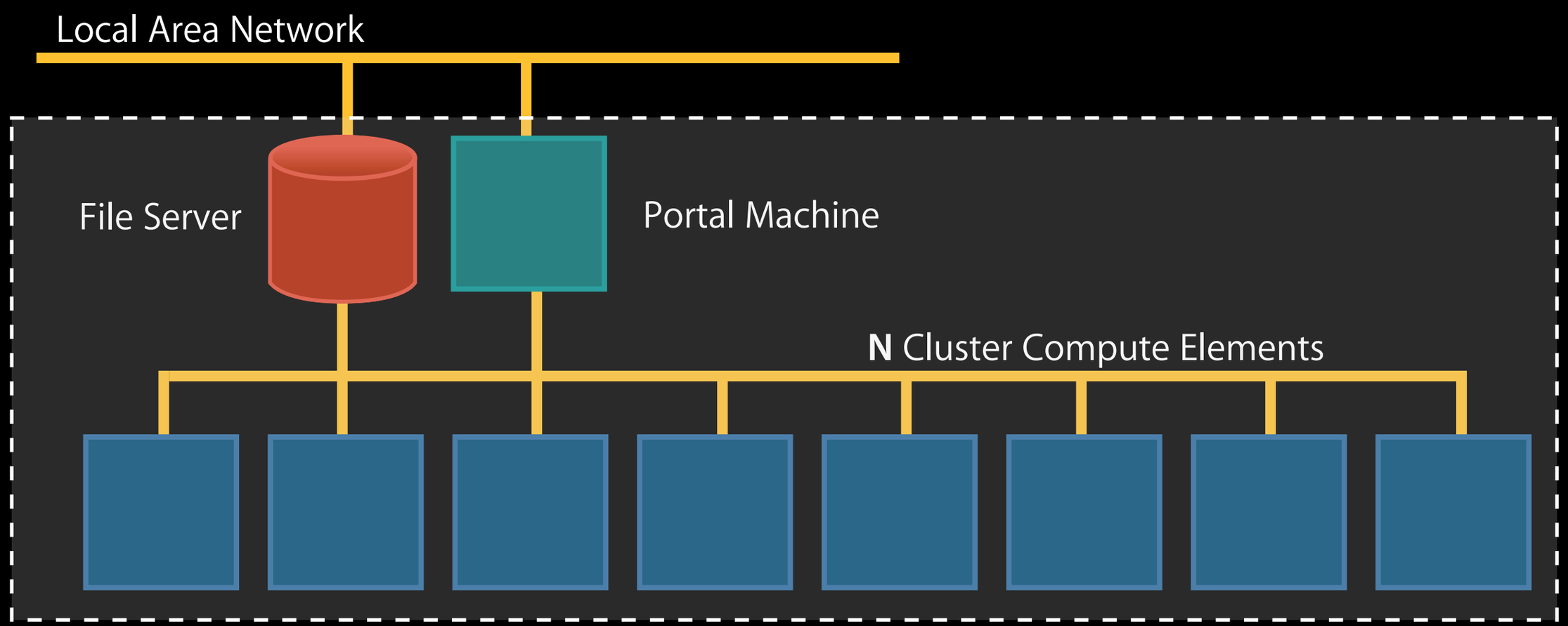
The BioTeam

Go From Many Computers…



The BioTeam

…to a Single Virtual Resource

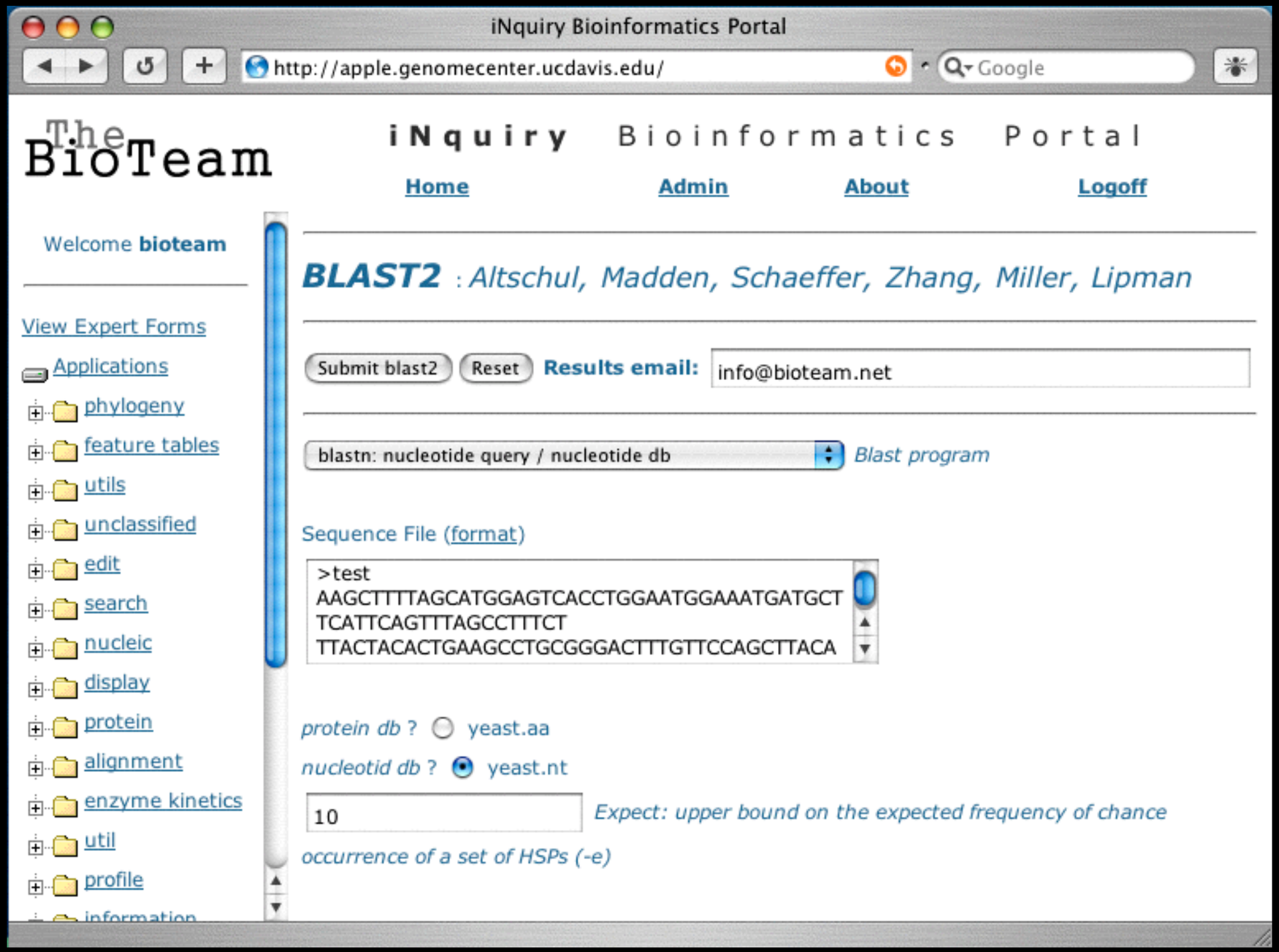


The BioTeam

From This…

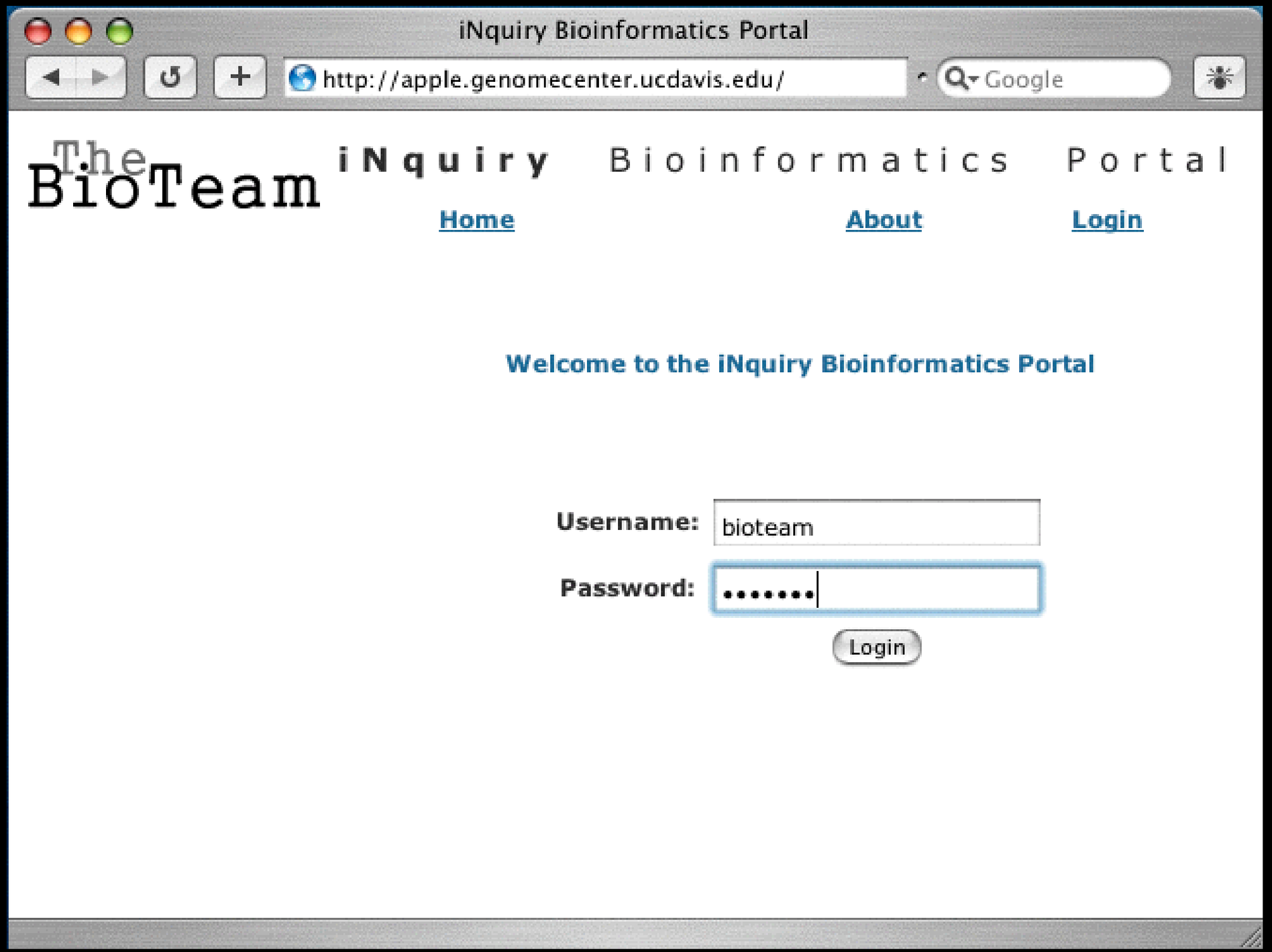
```
user% blastall -p blastn \
-i query.in -d /BlastDB/ \
yeast.nt -a 2 -o \
results.out
```

…to This



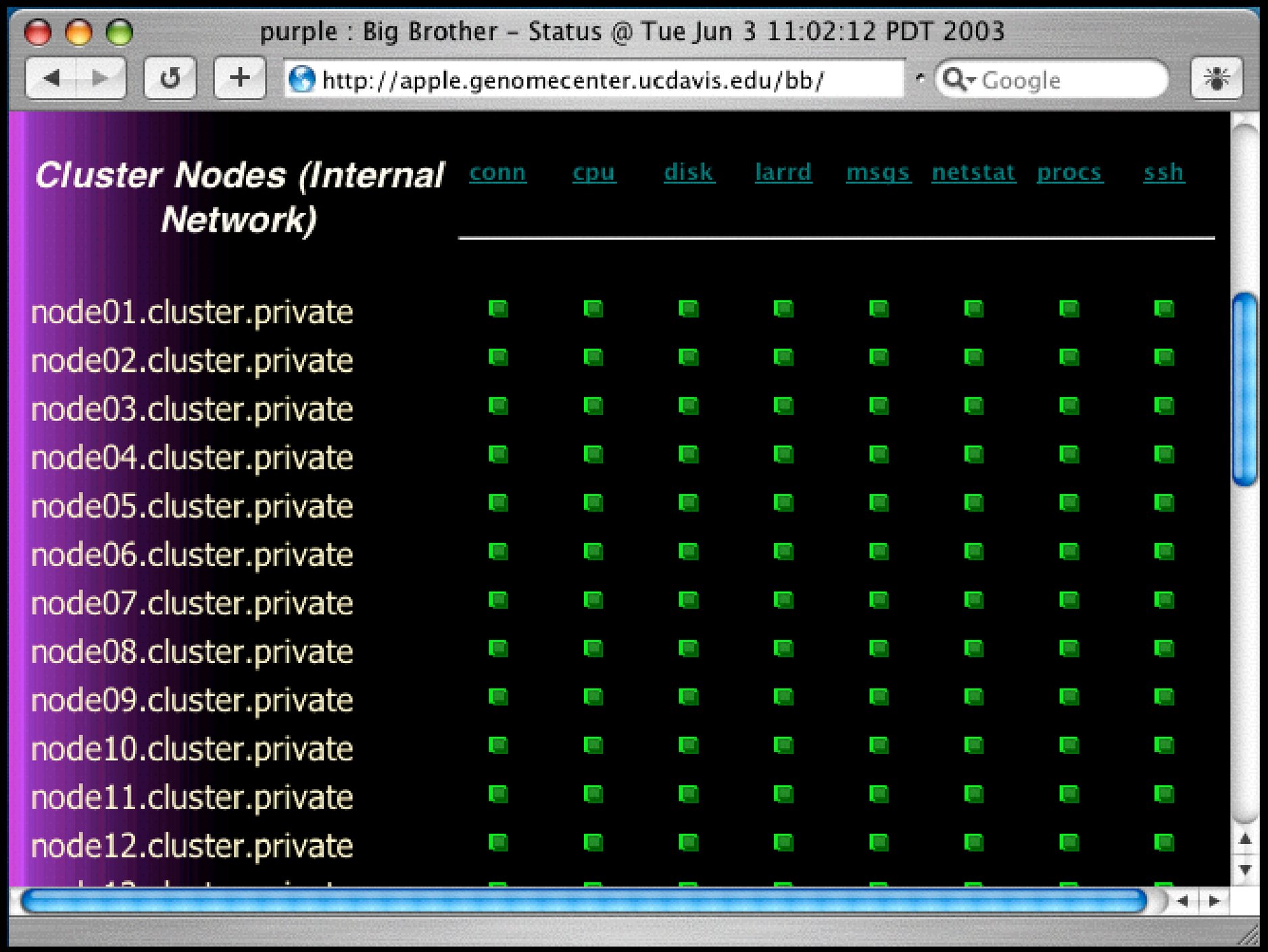
The BioTeam

Portal Interface



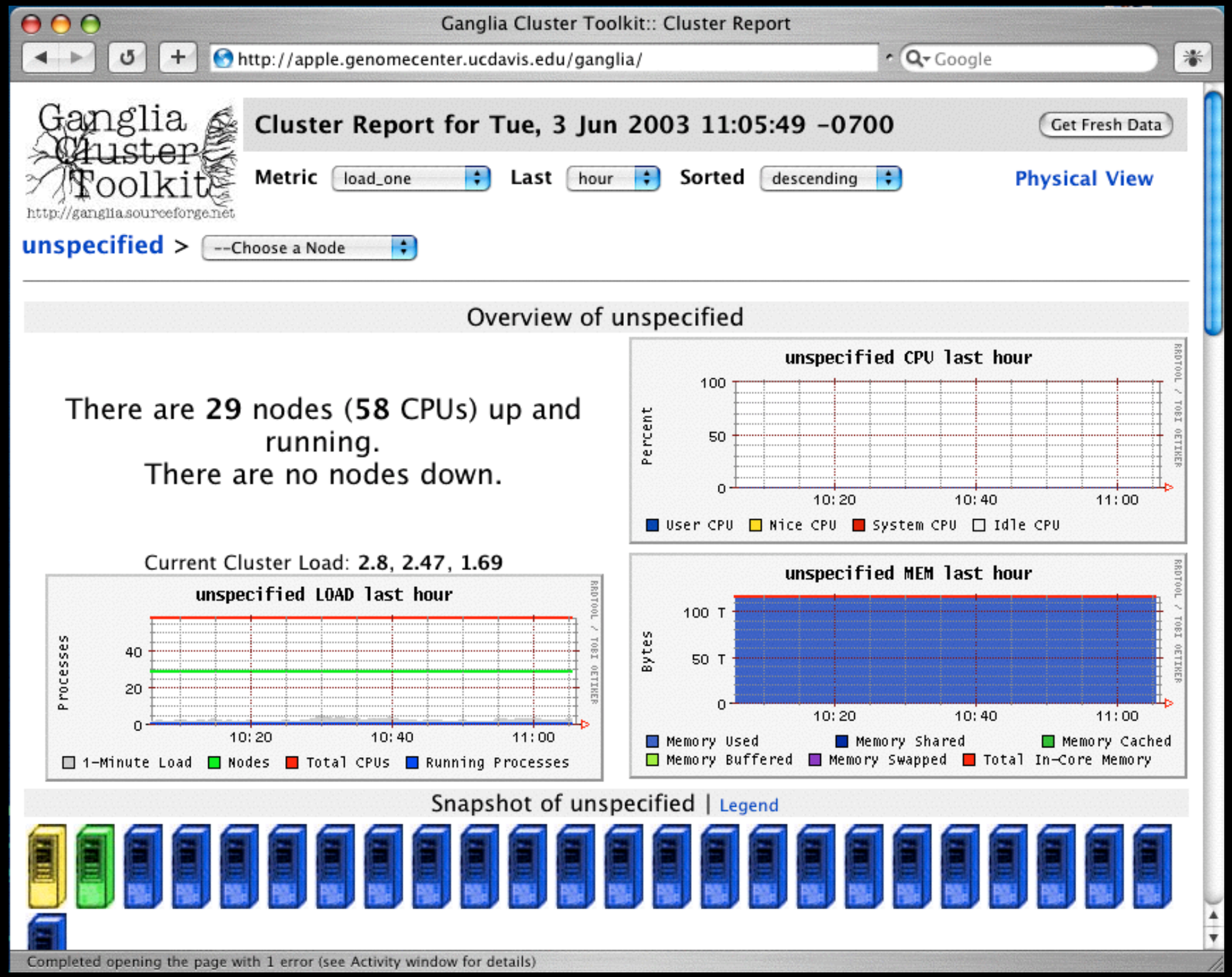
The BioTeam

Cluster Monitor—Big Brother



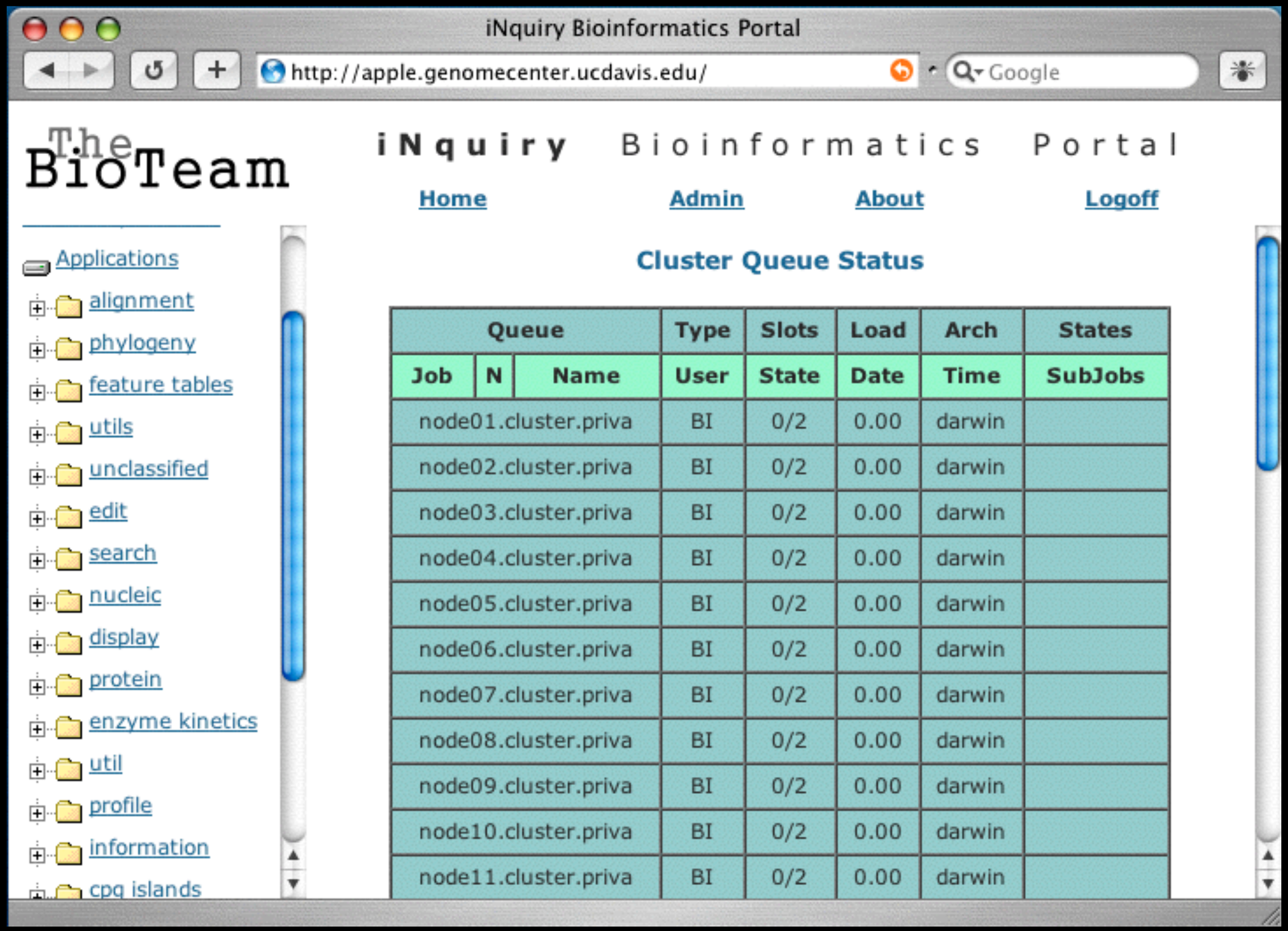
The BioTeam

Cluster Health—Ganglia



The BioTeam

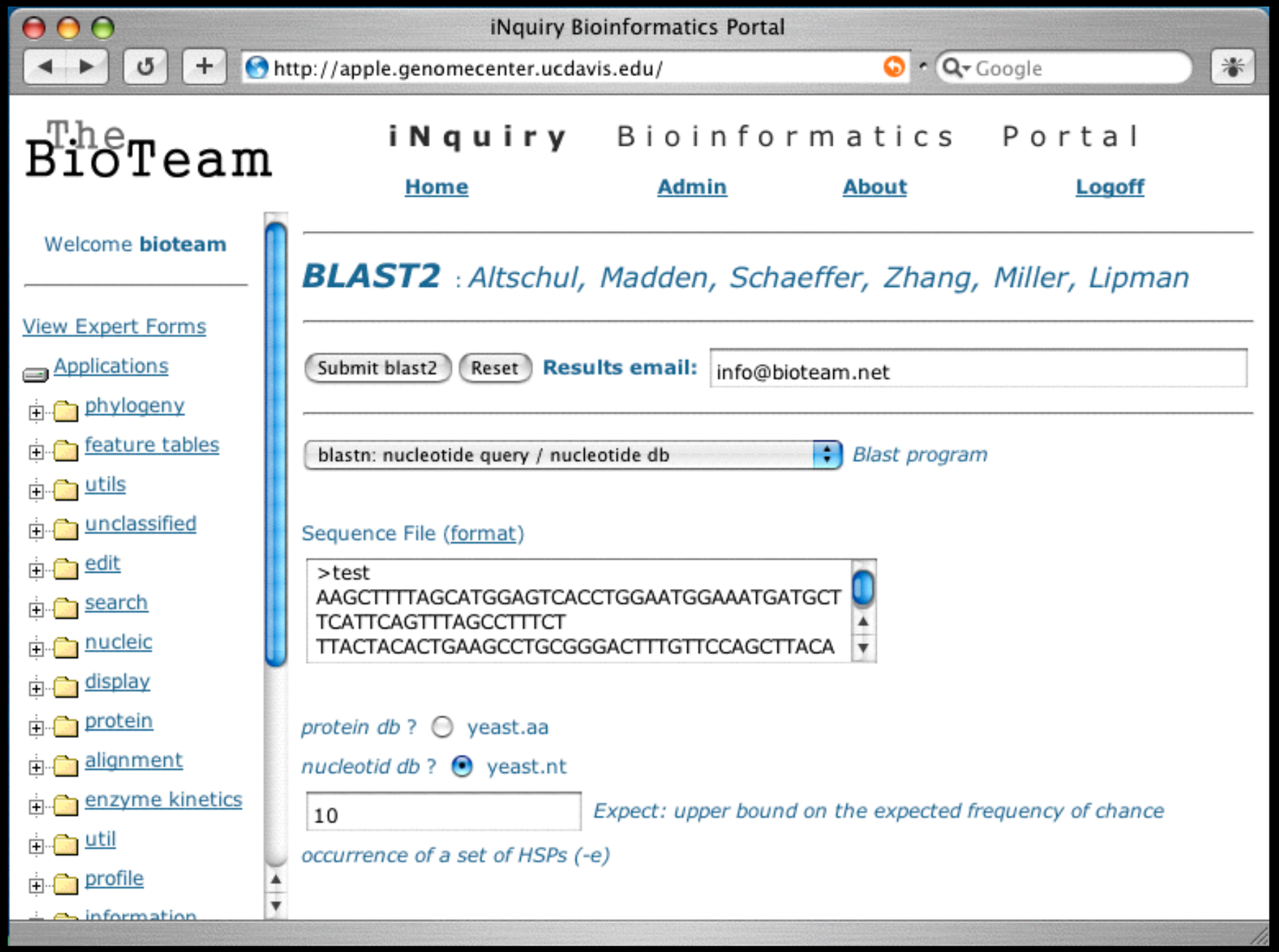
Cluster Jobs—SGE Monitor



The BioTeam

Consistent Application Interfaces

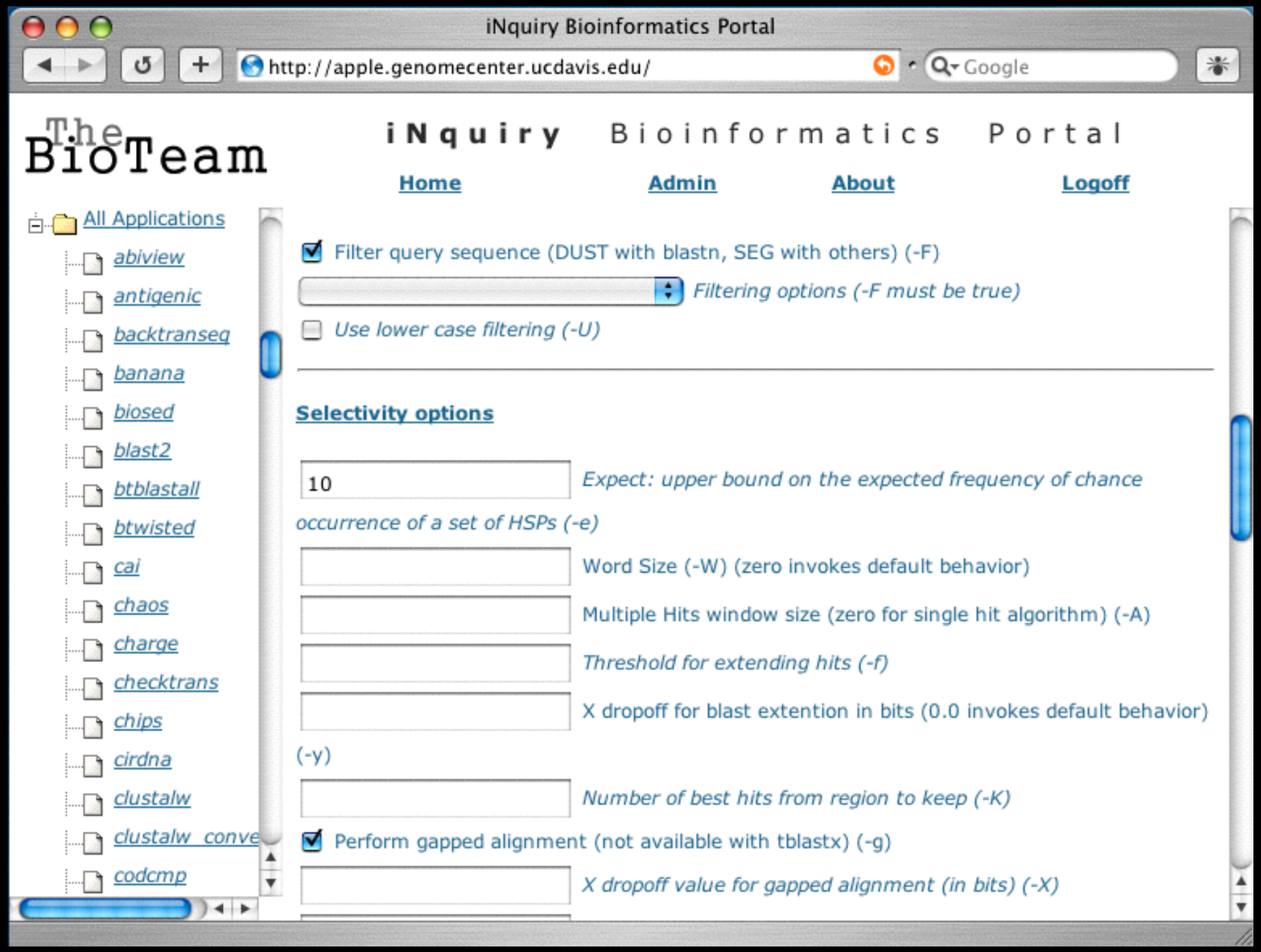
Simple View



The BioTeam

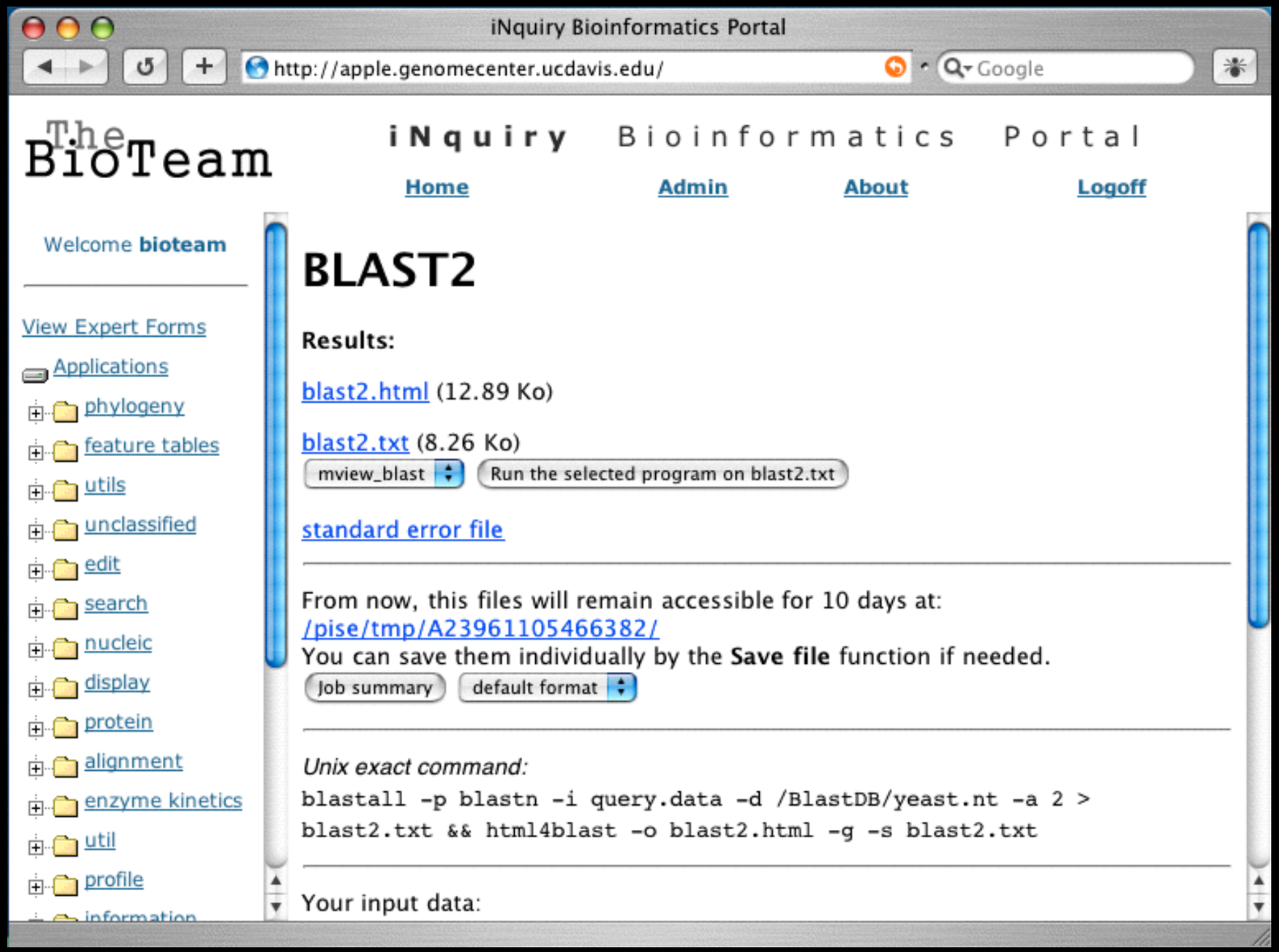
Consistent Application Interfaces

Expert View



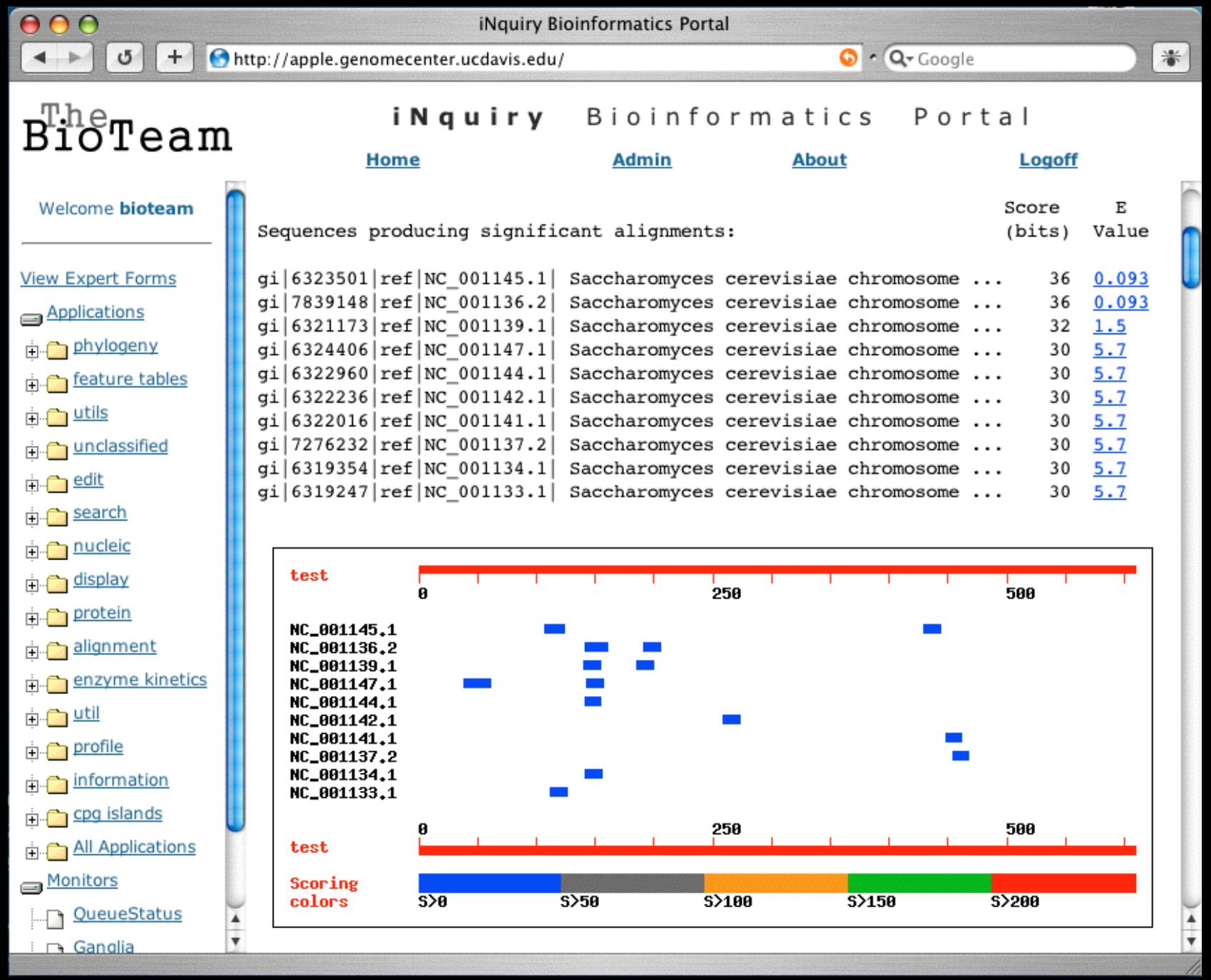
The BioTeam

Informatics Results



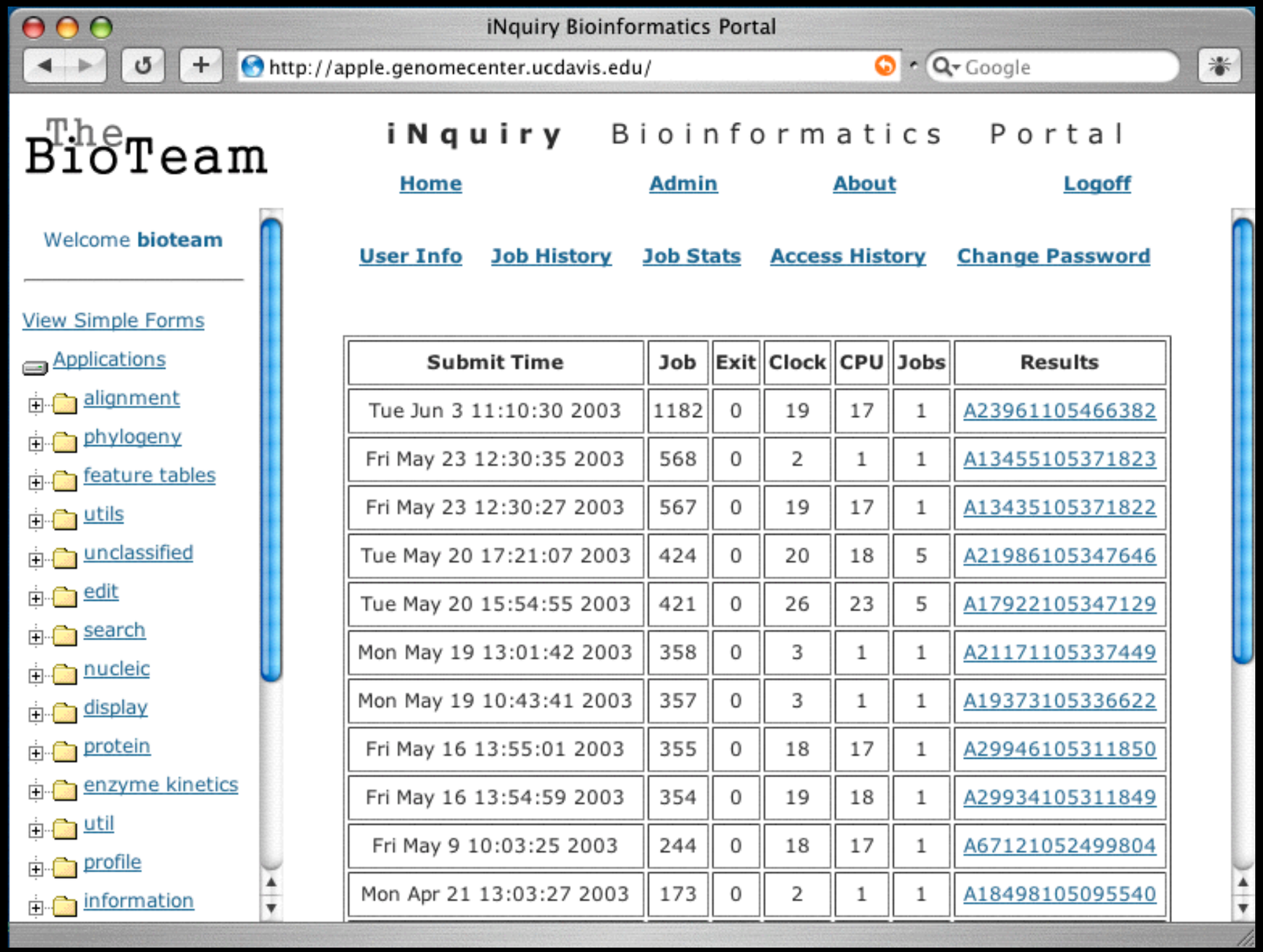
The BioTeam

Informatics Results



The BioTeam

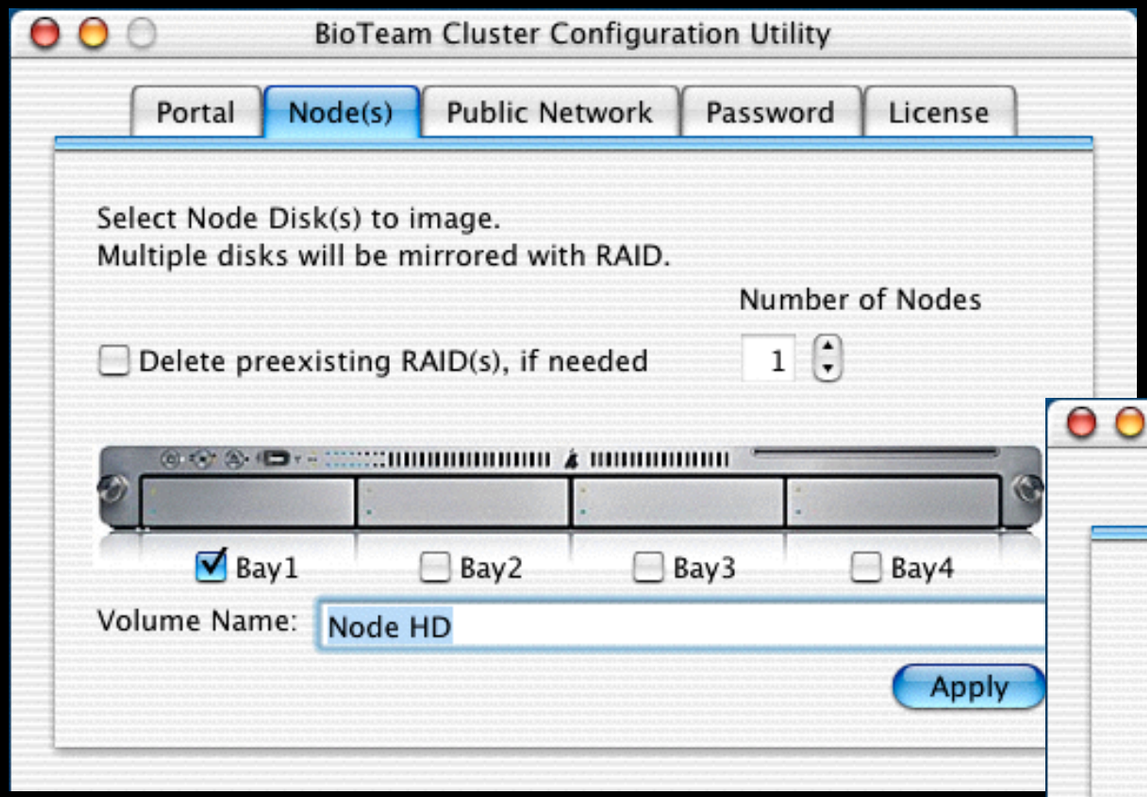
Experiment History

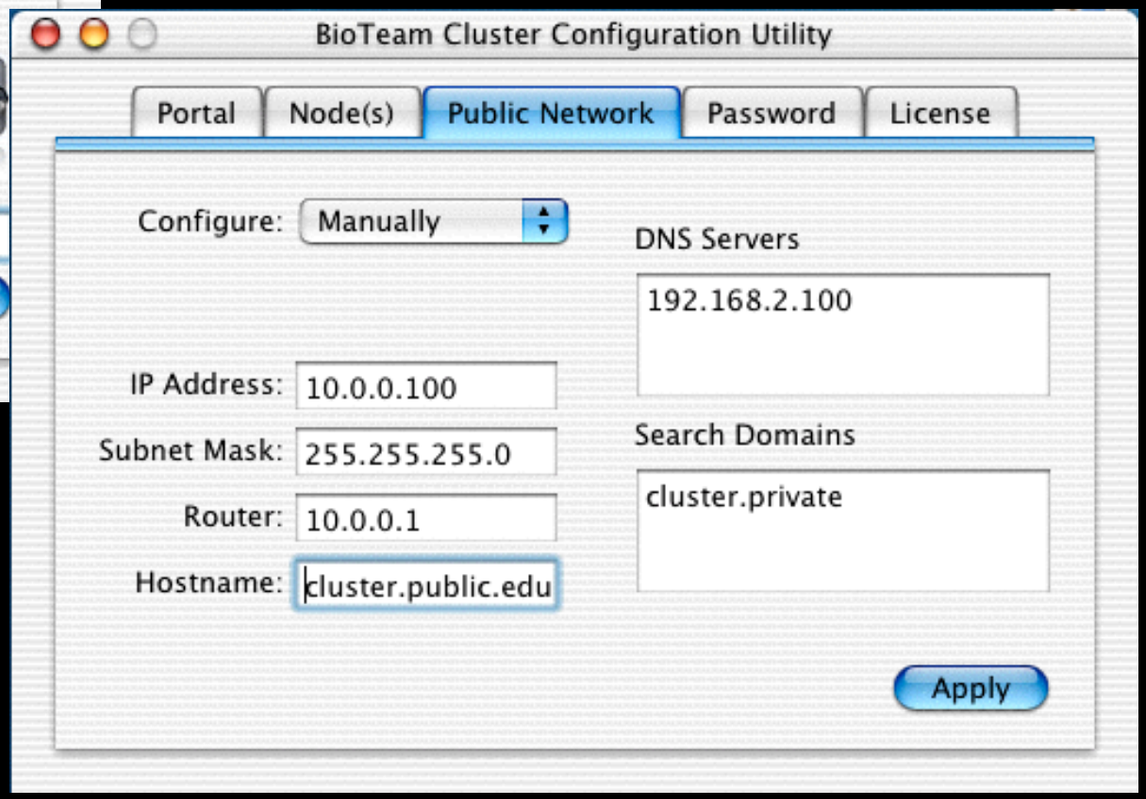


Submit Time	Job	Exit	Clock	CPU	Jobs	Results
Tue Jun 3 11:10:30 2003	1182	0	19	17	1	A23961105466382
Fri May 23 12:30:35 2003	568	0	2	1	1	A13455105371823
Fri May 23 12:30:27 2003	567	0	19	17	1	A13435105371822
Tue May 20 17:21:07 2003	424	0	20	18	5	A21986105347646
Tue May 20 15:54:55 2003	421	0	26	23	5	A17922105347129
Mon May 19 13:01:42 2003	358	0	3	1	1	A21171105337449
Mon May 19 10:43:41 2003	357	0	3	1	1	A19373105336622
Fri May 16 13:55:01 2003	355	0	18	17	1	A29946105311850
Fri May 16 13:54:59 2003	354	0	19	18	1	A29934105311849
Fri May 9 10:03:25 2003	244	0	18	17	1	A67121052499804
Mon Apr 21 13:03:27 2003	173	0	2	1	1	A18498105095540

iNquiry Installation

Step 1: Configure cluster with Aqua application

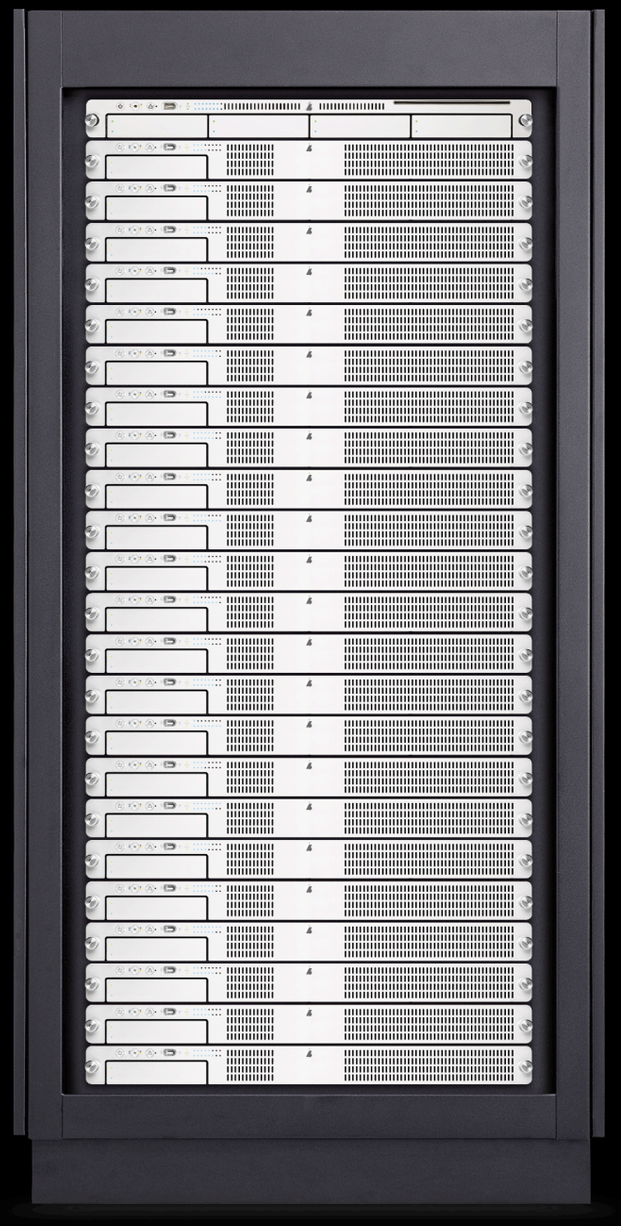




The BioTeam

iNquiry Installation

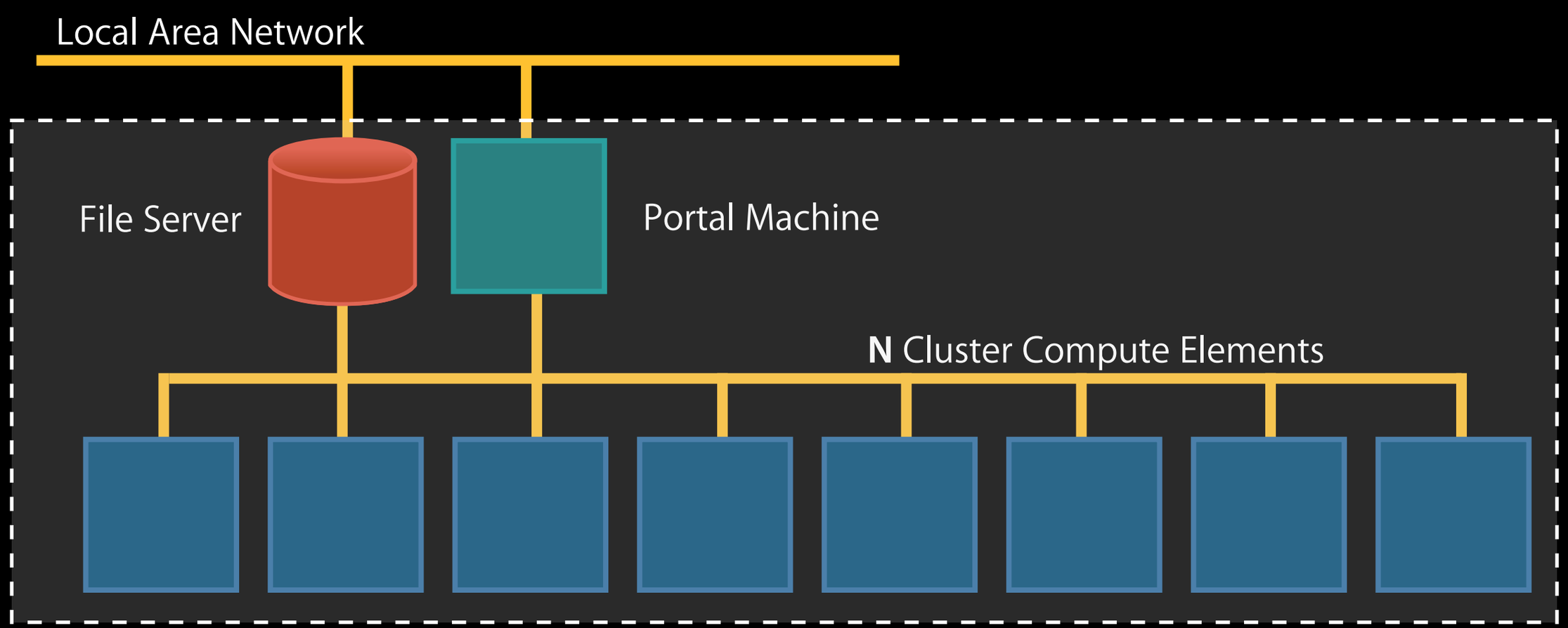
Step 2: Boot and Image Head Node from iPod



The BioTeam

iNquiry Installation

Step 3: NetBoot and Image Cluster Nodes from Head Node



The BioTeam

gridMathematica

Technical Computing for the Grid

Theodore Gray
Director of User Interfaces
gridmathematica@wolfram.com

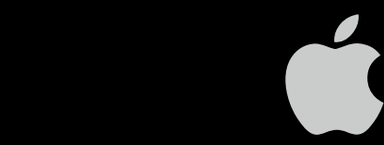
http://www.wolfram.com/products/gridmathematica/

gridMathematica

- Introduction
- Architecture
- Features
- Demo

Mathematica

- Fully integrated technical computing system
  - Powerful computation and visualization engine
  - Convenient, accessible user interface via Notebook interface
- Used by engineers, programmers, scientists, students
- Features
  - Symbolic and high-performance numeric computation
  - 2D and 3D data visualization
  - Programming interface
  - One-step creation of web documents
  - Mathematical typesetting and layout



gridMathematica

- Combines the power of Mathematica and modern clusters and grids to solve the most demanding problems in:
  - Mathematics
  - Science
  - Engineering
  - Finance

gridMathematica
Architecture

- High-level parallel commands
  - Remote evaluation
  - ParallelTable and ParallelMap
  - Other commands
  - Performance and efficiency
- Low-level concurrency commands
  - Basic concurrency constructs

gridMathematica

Features

- Distributed memory
- Master/slave parallelism
- Source code provided for all high-level commands
- Inter-kernel communications using MathLink
- Multiprocessor nodes or distribution across networks
- Scheduling of virtual processes or explicit distribution to available processors

gridMathematica
Features (Cont.)

- Virtual shared memory
- Synchronization, locking, and latency hiding
- Processes queues and concurrency
- Parallel functional programming
- Failure recovery and automatic reassignment of stranded processes

Demo

Wrap Up

Enterprise IT Roadmap

600 Apple Solutions in the Enterprise Learn how Apple's solutions fit into the Enterprise	**The Mission** **Monday, 3:30 pm**
604 Mac OS X Server Overview Mac OS X Server strategy and upcoming features	**The Mission** **Tuesday, 2:00 pm**
606 Mac OS X Server in Depth Learn the system capabilities of Mac OS X Server	**The Marina** **Tuesday, 3:30 pm**
621 Deploying Xserve What it takes to plan and manage deployment	**Russian Hill** **Thursday, 3:30 pm**
626 Deploying Xserve RAID What it takes to plan and manage deployment	**Pacific Heights** **Friday, 3:30 pm**

Developer Roadmap

301 Apple Tools for the UNIX Developer

Command line development capabilities of Mac OS X

Russian Hill
Tuesday, 9:00 am

305 Tuning Software with Performance Tools

Analyze, profile...and make it fast!

The Presido
Wednesday, 5:00 pm

Who to Contact

Douglas Brooks
Product Manager, Server Hardware, Apple
d.brooks@apple.com

Michael Athenas
Principal, The BioTeam
michael@bioteam.net

Theodore Gray
Director of User Interface, Wolfram
gridmathematica@wolfram.com

Skip Levens
Server Technology Evangelist
slevens@apple.com

http://developer.apple.com/wwdc2003/urls.html

For More Information

References

Cluster Resources

Xserve Compute Clustering
http://www.apple.com/server/clustering.html

Xserve, Xserve RAID and Mac OS X Server
http://www.apple.com/server

The BioTeam
http://www.bioteam.net

http://www.bioteam.net/MacOSX

gridMathematica
http://www.wolfram.com/products/gridmathematica/

References

Mailing Lists

SciTech List

http://www.lists.apple.com/mailman/listinfo/scitech/

UNIX Porting List

http://www.lists.apple.com/mailman/listinfo/unix-porting/

Biocluster List

http://bioinformatics.org/mailman/listinfo/bioclusters/

BioDarwin List

http://bioinformatics.org/mailman/listinfo/biodarwin/

Q&A

Douglas Brooks

Product Manager, Server Hardware

d.brooks@apple.com

http://developer.apple.com/wwdc2003/urls.html

WWDC 2003

Worldwide Developers Conference